DOCUMENT RESUME

ED 403 277 TM 025 951

AUTHOR Messick, Samuel
TITLE Validity and Washback in Language Testing.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-96-17
PUB DATE May 96
NOTE 21p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Applied Linguistics; *Construct Validity; Criteria; *Language Tests; Scores; *Simulation; Test Construction; Test Interpretation; Test Use; *Test Validity
IDENTIFIERS Authentic Assessment; *Authenticity; Direct Assessment; *Teaching to the Test; Testing Effects

ABSTRACT

The concept of "washback," especially prominent in the field of applied linguistics, refers to the extent to which a test influences teachers and learners to do things they would not otherwise necessarily do. Some writers invoke the notion of washback validity, holding that a test's validity should be gauged by the degree to which it has a positive influence on teaching. The complexity and uncontrolled variables of washback make it unsuitable for establishing test validity, but one can turn to the test properties likely to produce washback--authenticity and directness--and explore what they might mean in validity terms. The terms "authentic" and "direct" are most often used in connection with assessments involving realistic simulations or criterion samples. Purportedly authentic and direct performance assessments may not yield positive washback because the ideal forms of authenticity and directness rarely, if ever, exist. Construct underrepresentation and construct-irrelevant variance are present to varying degrees. To facilitate positive washback, an assessment must strive to avoid these two pitfalls. A comprehensive exploration of construct validity and its six distinguishable aspects (content, substantive, structural, generalizability, external, and consequential aspects) demonstrates that validity can be seen as a unified concept with the unifying force being the meaningfulness or interpretability of the test scores and action implications. The principles of unified validity provide a framework for evaluating all educational and psychological measurement, including washback. (Contains 29 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 403 277

RESEARCH REPORT

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

VALIDITY AND WASHBACK IN LANGUAGE TESTING

Samuel Messick

Educational Testing Service
Princeton, New Jersey
May 1996

TM025951

2

BEST COPY AVAILABLE

Copyright © 1996. Educational Testing Service. All rights reserved.

VALIDITY AND WASHBACK IN LANGUAGE TESTING

Samuel Messick[1]
Educational Testing Service

The current educational reform movement in the United States puts considerable stock in the notion that performance assessments, as opposed to multiple-choice tests, will facilitate improved teaching and learning (Resnick & Resnick, 1991; Wiggins, 1989, 1993). Some proponents even claim that performance assessments, especially those that are authentic and direct, are likely to be "systemically valid" in that they induce "in the education system curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure" (Frederiksen & Collins, 1989, p. 27).

A kindred notion prominent in applied linguistics, especially in Britain, is called "washback," which is the extent to which the test influences language teachers and learners to do things they would not otherwise necessarily do (Alderson & Wall, 1993). As with so-called systemic validity, some writers invoke the notion of "washback validity," holding that a test's validity should be gauged by the degree to which it has a positive influence on teaching (Morrow, 1986).

In the assessment of skills, tests having beneficial washback are likely to be criterion samples. That is, in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviors of listening, speaking, reading, and writing of the language being learned. Ideally, the move from learning exercises to test exercises should be seamless. As a consequence, for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing for the test.

Although only sparsely investigated to date, evidence of washback is typically sought in terms of behavioral and attitudinal changes in teachers and learners that are associated with the introduction of tests having

---

[1] For their stimulating and helpful comments on the manuscript, grateful acknowledgements are extended to Charles Alderson, Gary Buck, Gordon Hale, Ann Jungeblut, and Dianne Wall.

important educational consequences (Alderson & Wall, 1993). With respect to U.S. education reform, a more stringent claim has been made involving not only changes in teacher and learner behaviors but also in learner outcomes. To wit, "evidence for systemic validity would be an improvement in [the tested] skills after the test has been in place within the educational system for a period of time" (Frederiksen & Collins, 1989, p. 27).

However, such forms of evidence are only circumstantial with respect to test validity in that a poor test may be associated with positive effects and a good test with negative effects because of other things that are done or not done in the educational system. Technically speaking, such effects should not be viewed as test washback but rather as due to good or bad educational practices apart from the quality of the test. Furthermore, a test might influence ***what*** is taught but not ***how*** it is taught, might influence ***teacher*** behaviors but not ***learner*** behaviors, or might influence both with little or no improvement in skills. Hence, washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational scene. Indeed, if it exists, washback "is likely to be a complex phenomenon which cannot be related directly to a test's validity" (Alderson & Wall, 1993, p. 116). In any event, washback is only one form of testing consequence that needs to be weighed in evaluating validity, and testing consequences are only one aspect of construct validity needing to be addressed. Neither testing consequences in general nor washback in particular can stand alone as a standard of validity.

Hence, one should not rely on washback, with all its complexity and uncontrolled variables, to establish test validity, Morrow (1986) and Frederiksen and Collins (1989) notwithstanding. Rather, one can instead turn to the test properties likely to produce washback -- namely, authenticity and directness -- and ask what they might mean in validity terms. Next, we examine the implications of authenticity and directness for test validity and then cast the issues in the broader context of a comprehensive view of construct validity.

The broader concept is emphasized for two main reasons. First, in this validity framework, washback is seen as an instance of the consequential aspect of construct validity, which, along with five other important aspects,

address the key questions that need to be answered in evaluating test validity. Second, by focussing not on washback per se but on the deeper and more encompassing issue of validity, we highlight the multiple forms of evidence needed to sustain valid language test use. In particular, by attempting to minimize sources of invalidity in language test design, the test deficiencies and contaminants that stimulate negative washback are also minimized, thereby increasing the likelihood of positive washback.

In short, we emphasize first the need to establish valid evidential grounds for trustworthy inferences about tested language proficiency to provide a basis for distinguishing test-linked positive washback from good teaching regardless of the quality of the test and negative washback from poor teaching. This is important because, technically speaking, evidence of teaching and learning effects should be interpreted as washback -- either in general or in particular as contributing to the consequential aspect of construct validity -- only if that evidence can be linked to the introduction and use of the test.

AUTHENTICITY AND DIRECTNESS AS VALIDITY STANDARDS

The two terms "authentic" and "direct" are most often used in connection with assessments involving realistic simulations or criterion samples. Because it is widely thought in some educational circles that authenticity and directness of assessment facilitate positive consequences for teaching and learning (e.g., Resnick & Resnick, 1991; Wiggins, 1993), they constitute tacit validity standards, so we need to address what these labels might mean in validity terms.

Minimizing Sources of Invalidity

Ideally, ***authentic*** assessments pose engaging and worthy tasks (usually involving multiple processes) in realistic settings or close simulations so that the tasks and processes, as well as available time and resources, parallel those in the real world. The major measurement concern of authenticity is that nothing important be left out of the assessment of the focal construct (Messick, 1994). This is tantamount to the general validity

standard of minimal construct underrepresentation. However, although authenticity implies minimal construct underrepresentation, the obverse does not hold. This is the case because minimal construct underrepresentation does not necessarily imply the close simulation of real-world processes and resources typically associated with authenticity in the current educational literature on performance assessment.

Ideally, *direct* assessments involve open-ended tasks in which the respondent can freely perform the complex skill at issue unfettered by structured item forms or restrictive response formats. The intent is to minimize constraints on examinee behavior associated with sources of construct-irrelevant method variance such as testwiseness in coping with various item-types, differential tendencies toward guessing, and other artificial restrictions on examinees' representations of problems and on their modes of thinking or response. Thus, the major measurement concern of directness is that nothing irrelevant be added that interferes with or contaminates construct assessment. This is tantamount to the general validity standard of minimal construct-irrelevant variance (Messick, 1994). Incidentally, the term "direct assessment" is a misnomer because it always promises too much. In education and psychology, "all measurements are indirect in one sense or another" (Guilford, 1936, p. 5). Measurement always involves, even if only tacitly, intervening processes of judgment, comparison, or inference.

In the threat to validity known as *construct underrepresentation* (which jeopardizes authenticity), the assessment is deficient: The test is too narrow and fails to include important dimensions or facets of focal constructs. In the threat to validity known as *construct-irrelevant variance* (which jeopardizes directness), the assessment is too broad, containing excess reliable variance that is irrelevant to the interpreted construct. Both threats are operative in all assessments. However, as always in test validation, the critical issue is the gathering of sufficiently compelling evidence to counter these two major threats to construct validity.

A comprehensive unified view of construct validity will be considered shortly as a means of addressing an interrelated set of perennial validity questions. But first let us briefly examine the widely anticipated connection

of authentic and direct assessments with washback, a link that Alderson and Wall (1990) maintain is still evidentially tenuous at best.

Facilitating Positive Washback

There are a number of reasons why purportedly authentic and direct performance assessments do not readily yield positive washback. Some reasons pertain to properties of the assessment itself and others to properties of the educational system, especially of the instructional and assessment setting.

To begin with, the ideal forms of authenticity and directness rarely if ever exist. To some degree, construct underrepresentation and construct-irrelevant variance are ever with us. The test is never a completely faithful exemplar of criterion behaviors. This is so for at least two reasons. First, by its very nature, the test is likely to evoke evaluative anxiety and attendant coping processes that are not operative in the criterion performance, at least not in the same way (Loevinger, 1957). Second, the test performances are scored and interpreted in ways that are unlikely to fully or faithfully capture the criterion domain processes. In language testing, as in all educational and psychological measurement, what matters are not the processes operative in task performance, exemplary though they may be, but the processes captured in test scoring and interpretation. If it occurs, washback is likely to be oriented toward the achievement of high test scores as opposed to the attainment of facile domain skills. Thus, to facilitate positive washback, the assessment must strive to minimize construct underrepresentation and construct-irrelevant difficulty in the interpreted scores.

With respect to the instructional and assessment setting, there are a number of links in the chain that ostensibly binds the test to positive washback, and these links need to be more strongly forged than is ordinarily the case. Specifically, for performance assessments to "fulfill their promise of driving improvements in student learning and achievement, assessment systems must incorporate the means for affecting what teachers do and how they think about what they do in their classrooms" (Sheingold, Heller, & Paulukonis, 1995). For example, in one effort to achieve systemic validity or positive washback, the assessment system involved teachers responsible for defining, creating, and revising the assessment tasks, with the aid of cognitive supports in the form of guiding questions and design guidelines as

well as the social support of the teachers working collaboratively among themselves and with outside experts. Thus, washback appears to depend on a number of important factors in the educational system in addition to the validity of the tests.

We turn now to a comprehensive view of construct validity as a means of integrating complementary forms of evidence pertinent to validity, including evidence of washback.

COMPREHENSIVENESS OF CONSTRUCT VALIDITY

Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the ***adequacy*** and ***appropriateness*** of ***interpretations*** and ***actions*** based on test scores or other modes of assessment (Messick, 1989). Validity is not a property of the test or assessment as such, but rather of the meaning of the test scores. Hence, what is to be validated is not the test or observation device per se but rather the inferences derived from test scores or other indicators (Cronbach, 1971) -- inferences about score meaning or interpretation and about the implications for action that the interpretation entails.

For example, a validated proficiency test can be subverted by coaching, or test preparation practices emphasizing testwiseness strategies, that might increase test scores without correspondingly improving the skills measured by the test. Although this would not compromise the validity of the uncoached test in general, the validity of the interpretation and use of the coached scores would be jeopardized. In contrast, test preparation practices emphasizing test familiarization and anxiety reduction may actually improve validity: Scores that formerly were invalidly low because of anxiety might now become validly higher (Messick, 1982).

In essence, then, test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use. Because score meaning is a ***construction*** that makes theoretical sense out of both the performance regularities summarized by the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as ***construct*** validity.

Perennial Validity Questions

To evaluate the meaning and consequences of measurement is no small order, however, and requires attention to a number of persistent validity questions, such as:

- Are we looking at the right things in the right balance?
- Has anything important been left out?
- Does our way of looking introduce sources of invalidity or irrelevant variance that bias the scores or judgments?
- Does our way of scoring reflect the manner in which domain processes combine to produce effects and is our score structure consistent with the structure of the domain about which inferences are to be drawn or predictions made?
- What evidence is there that our scores mean what we interpret them to mean, in particular, as reflections of personal attributes or competencies having plausible implications for educational action?
- Are there plausible rival interpretations of score meaning or alternative implications for action and, if so, by what evidence and arguments are they discounted?
- Are the judgments or scores reliable and are their properties and relationships generalizable across the contents and contexts of use as well as across pertinent population groups?
- Are the value implications of score interpretations empirically grounded, especially if pejorative in tone, and are they commensurate with the score's trait implications?
- Do the scores have utility for the proposed purposes in the applied settings?
- Are the scores applied fairly for these purposes, that is, consistently and equitably across individuals and groups?
- Are the short- and long-term consequences of score interpretation and use supportive of the general testing aims and are there any adverse side-effects?

Which, if any, of these questions is unnecessary to address in justifying score interpretation and use? Which, if any, can be forgone in validating the interpretation and use of performance assessments or other modes of assessment? The general thrust of such questions is to seek evidence

and arguments to discount the two major threats to construct validity -- namely, construct underrepresentation and construct-irrelevant variance -- as well as to evaluate the action implications of score meaning.

Addressing these questions with solid evidence is important both in general to justify test use and in particular in connection with the current emphasis on washback. For example, attempting to improve validity by test design, as is implied by many of these questions, may increase the likelihood of positive washback. In turn, evidence of washback contributes to the consequential aspect of construct validity. Furthermore, information about the operative level of test validity should help one distinguish test washback per se from the effects of good or bad educational practices regardless of the quality of the test.

With regard to the latter point, if a test's validity is compromised because of construct underrepresentation or construct-irrelevant variance, it is likely that any signs of good teaching or learning associated with the use of the test are only circumstantial and more likely due to good educational practices regardless of test use. Similarly, signs of poor teaching or learning associated with the use of a construct-validated test are more likely to reflect poor educational practices regardless of test use. That is, it is problematic to claim evidence of test washback if a logical or evidential link cannot be forged between the teaching or learning outcomes and the test properties thought to influence them. Although there may be exceptions requiring careful scrutiny, negative washback per se should be associated with the introduction and use of less valid tests and positive washback with the introduction and use of more valid tests because construct underrepresentation and construct-irrelevant variance in the test could precipitate bad educational practices while minimizing these threats to validity should facilitate good educational practices.

Aspects of Construct Validity

Although validity is now widely viewed as an integral or unified concept (APA, 1985), this does not imply answering only one overarching validity question or even several questions separately or one at a time. Rather, it implies an integration of multiple complementary forms of convergent and discriminant evidence to answer an interdependent set of questions such as

those in the previous section. To make this explicit, it is illuminating to differentiate unified validity into several distinct aspects to underscore issues and nuances that might otherwise be downplayed or overlooked, such as the social consequences of performance assessments or the role of score meaning in applied use.

In particular, six distinguishable aspects of construct validity are highlighted as a means of addressing central issues implicit in the notion of validity as a unified concept. These are content, substantive, structural, generalizability, external, and consequential aspects of construct validity. In effect, these six aspects function as general validity criteria or standards for all educational and psychological measurement (Messick, 1989). They are briefly characterized as follows:

- The content aspect of construct validity (Lennon, 1956; Messick, 1989) includes evidence of content relevance and representativeness as well as of technical quality (e.g., appropriate reading level, unambiguous phrasing, and correct keying).

- The substantive aspect refers to theoretical rationales for the observed performance regularities and item correlations, including process models of task performance (Embretson, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.

- The structural aspect (Loevinger, 1957; Messick, 1989) appraises the fidelity of the score scales to the structure of the construct domain at issue with respect to both number (i.e., appropriate dimensionality) and makeup (e.g., conjunctive vs. disjunctive, trait vs. class).

- The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Feldt & Brennan, 1989; Shulman, 1970), including generalizability of test-criterion relationships across settings and time periods, which is known as "validity generalization" (Hunter, Schmidt, & Jackson, 1982).

- The external aspect includes convergent and discriminant evidence from multitrait-multimethod comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Cronbach & Gleser, 1965).

- The consequential aspect appraises the value implications of score interpretation as a basis for action as well as the actual and

potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989), as well as to washback.

A key issue for the ***content*** aspect of construct validity is the specification of the boundaries of the construct domain to be assessed -- that is, determining the knowledge, skills, and other attributes to be revealed by the assessment tasks. The boundaries and structure of the construct domain can be addressed by means of job analysis, task analysis, curriculum analysis, and especially domain theory, that is, scientific inquiry into the nature of the domain processes and the ways in which they combine to produce effects or outcomes. A major goal of domain theory is to understand the construct-relevant sources of task difficulty, which then serves as a guide to the rational development and scoring of performance tasks. At whatever stage of its development, then, domain theory is a primary basis for specifying the boundaries and structure of the construct to be assessed.

However, it is not sufficient merely to select tasks that are relevant to the construct domain. In addition, the assessment should assemble tasks that are representative of the domain in some sense. The intent is to insure that all important parts of the construct domain are covered, which is usually described as selecting tasks that sample domain processes in terms of their functional importance. Both the content relevance and representativeness of assessment tasks are traditionally appraised by expert professional judgment, documentation of which serves to address the content aspect of construct validity.

The ***substantive*** aspect of construct validity emphasizes two important points: One is the need for tasks providing appropriate sampling of domain ***processes*** in addition to traditional coverage of domain ***content***; the other is the need to move beyond traditional professional judgment of content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance. Thus, the substantive aspect adds to the content aspect of construct validity the need for empirical evidence of response consistencies or performance regularities reflective of domain processes (Embretson, 1983; Loevinger, 1957; Messick, 1989).

According to the ***structural*** aspect of construct validity, scoring models should be rationally consistent with what is known about the structural relations inherent in behavioral manifestations of the construct in question (Loevinger, 1957; Peak, 1953). That is, the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria and rubrics. Thus, the internal structure of the assessment (i.e., interrelations among the scored aspects of task and subtask performance) should be consistent with what is known about the internal structure of the construct domain (Messick, 1989).

In the ***generalizability*** aspect of construct validity, the concern is that a performance assessment should provide representative coverage of the content and processes of the construct domain. This is meant to insure that the score interpretation not be limited to the sample of assessed tasks but be generalizable to the construct domain more broadly. Evidence of such generalizability depends on the degree of correlation of the assessed tasks with other tasks representing the construct or aspects of the construct. This issue of generalizability of score inferences across tasks and contexts goes to the very heart of score meaning. Indeed, setting the boundaries of score meaning is precisely what generalizability evidence is meant to address.

The emphasis here is on generalizability in two senses, namely, as it bears on reliability and on transfer. Generalizability as reliability (Feldt & Brennan, 1989) refers to the consistency of performance across the tasks, occasions, and raters of a particular assessment, which might be quite limited in scope. For example, we have all been concerned that some assessments with a narrow set of tasks might attain higher reliability in the form of cross-task consistency, but at the expense of construct validity. In contrast, generalizability as transfer requires consistency of performance across tasks that are representative of the broader construct domain. That is, transfer refers to the range of tasks that performance on the assessed tasks facilitates the learning of or, more generally, is predictive of (Ferguson, 1956). Thus, generalizability evidence becomes especially pertinent to washback if learning in preparation for the test tasks facilitates the learning of an array of related tasks to improve or solidify domain proficiency. Generalizability as transfer depends not only on

generalizability theory but also on construct theory. In essence, then, generalizability evidence is an aspect of construct validity because it establishes boundaries on the meaning of the construct scores.

However, because of the extensive time required for the typical performance task, there is a conflict in performance assessment between time-intensive depth of examination and the breadth of domain coverage needed for generalizability of construct interpretation. This conflict between depth and breadth of coverage is often viewed as entailing a trade-off between validity and reliability (or generalizability). It might better be depicted as a trade-off between the valid description of the specifics of a complex task performance and the power of construct interpretation. In any event, such a conflict signals a design problem that needs to be carefully negotiated in performance assessment (Wiggins, 1993).

The ***external*** aspect of construct validity refers to the extent to which the assessment scores' relationships with other measures and nonassessment behaviors reflect the expected high, low, and interactive relations implicit in the theory of the construct being assessed. Thus, the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning. That is, the constructs represented in the assessment should rationally account for the external pattern of correlations.

Of special importance among these external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, certification of competence, program evaluation, or other accountability purposes in applied settings. Once again, the construct theory points to the relevance of potential relationships between the assessment scores and criterion measures, and empirical evidence of such links attests to the utility of the scores for the applied purpose.

The ***consequential*** aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term, especially those associated with bias in scoring and interpretation, with unfairness in test use, and with positive or negative washback effects on teaching and learning. However, this form of evidence should not be viewed in isolation as a separate type of validity, say, of "consequential validity" or, worse still, "washback

validity." Rather, because the social values served in the intended and unintended outcomes of test interpretation and use both derive from and contribute to the meaning of the test scores, appraisal of social consequences of the testing is also seen to be subsumed as an aspect of construct validity (Messick, 1980).

Consequences associated with testing are likely to be a function of numerous factors in the context or setting and in the persons responding as well as in the content and form of the test. To bear on validity, convergent and discriminant evidence should be accrued linking positive washback or any positive consequences to the test use. This might be accomplished, for example, by means of classroom observations or questionnaires documenting changes in teacher and learner behavior associated with the introduction of the test.

The primary measurement concern with respect to adverse consequences is that negative washback or, indeed, any negative impact on individuals or groups should not derive from any source of test invalidity such as construct underrepresentation or construct-irrelevant variance (Messick, 1989). That is, invalidly low scores should not occur because the assessment is missing something relevant to the focal construct that, if present, would have permitted the affected persons to display their competence. Moreover, invalidly low scores should not occur because the measurement contains something irrelevant that interferes with the affected persons' demonstration of competence.

Furthermore, if what is underrepresented in the assessment of communicative competence is an important part of the criterion performance, such as listening and speaking as opposed to reading and writing, then invalidly high scores may be attained by examinees well-prepared on the represented skills but ill-prepared on the underrepresented ones. That is, scores may be invalidly high as indicators of communicative competence even though they are valid measures of reading and writing proficiency. Invalidly high scores may also be obtained by testwise examinees who are facile in dealing with construct-irrelevant difficulty.

It would seem, then, that if one is concerned with fostering positive washback and reducing negative washback, one should concentrate first on minimizing construct underrepresentation and construct-irrelevant difficulty

in the assessment. That is, ***rather than seeking washback as a sign of test validity, seek validity by design as a likely basis for washback***. To accomplish this, one need not insist on assessments that are realistic or authentic and open-ended or direct. Pragmatically, the touchstone is an assessment that adequately represents the focal construct using formats that are acceptably obtrusive within the practical constraints of feasible test administration and scoring, that is, formats in which method variance is relatively minor and can be taken into account in scoring and interpretation.

In practice, testmakers are mainly concerned about adverse consequences that are traceable to sources of test invalidity such as construct underrepresentation and construct-irrelevant difficulty. These concerns are especially salient in connection with issues of bias, fairness, and distributive justice, but also potentially with respect to negative washback. For example, if important constructs or aspects of constructs are underrepresented on the test, teachers might come to overemphasize those constructs that are well-represented and downplay those that are not. If the test employs unfamiliar item formats or stresses knowledge of grammar, for instance, to the detriment of communicative competence, teachers might pay undue attention to overcoming the irrelevant difficulty as opposed to fostering communicative proficiency. One defense against such adverse consequences is to provide test familiarization and preparation materials to reduce the effects of construct-irrelevant difficulty and attendant test anxiety, but the best defense is to minimize such irrelevant difficulty in the first place as well as construct underrepresentation.

In contrast, adverse consequences associated with the valid measurement of current status -- such as validly low scores resulting from poor teaching or limited opportunity to learn -- are not the testmakers' responsibility. Such adverse consequences of valid assessment represent problems not of measurement, but rather of teaching and of educational or social policy.

From the discussion thus far, it should be clear that test validity cannot ***rely*** on any one of the complementary forms of evidence just discussed. However, neither does validity ***require*** any one form, granted that there is defensible convergent and discriminant evidence supporting score meaning. To the extent that some form of evidence cannot be developed -- as when criterion-related studies must be forgone because of small sample sizes,

unreliable or contaminated criteria, and highly restricted score ranges -- heightened emphasis can be placed on other evidence, especially on the construct validity of the predictor tests and the relevance of the construct to the criterion domain (Guion, 1976; Messick, 1989). What *is* required is a compelling argument that the available evidence justifies the test interpretation and use, even though some pertinent evidence had to be forgone. Hence, validity becomes a unified concept and the unifying force is the meaningfulness or trustworthy interpretability of the test scores and their action implications, namely, construct validity.

Validity As Integrative Summary

The six aspects of construct validity apply to all educational and psychological measurement, including performance assessments or other alternative assessment modes. Taken together, they provide a way of addressing the multiple and interrelated validity questions that need to be answered in justifying score interpretation and use. In previous writings I maintained that it is "the relation between the evidence and the inferences drawn that should determine the validation focus" (Messick, 1989. p. 16). This relation is embodied in theoretical rationales or persuasive arguments that the obtained evidence both supports the preferred inferences and undercuts plausible rival inferences. From this perspective, as Cronbach (1988) concluded, validation is evaluation argument. That is, as stipulated earlier, validation is empirical evaluation of the meaning and consequences of measurement. The term "empirical evaluation" is meant to convey that the validation process is scientific as well as rhetorical and requires both evidence and argument.

By focussing on the argument or rationale employed to support the assumptions and inferences invoked in the score-based interpretations and actions of a particular test use, one can prioritize the forms of validity evidence needed in terms of the important points in the argument that require justification or support (Kane, 1992; Shepard, 1993). Helpful as this may be, there still remain problems in setting priorities for needed evidence because the argument may be incomplete or off target, not all the assumptions may be addressed, and the need to discount alternative arguments evokes multiple priorities.

The point here is that the six aspects of construct validity afford a means of checking that the theoretical rationale or persuasive argument linking the evidence to the inferences drawn touches the important bases and, if not, requiring that an argument be provided that such omissions are defensible. They are highlighted because most score-based interpretations and action inferences, as well as the elaborated rationales or arguments that attempt to legitimize them (Kane, 1992), either invoke these properties or assume them, explicitly or tacitly.

That is, most score interpretations refer to relevant content and operative processes, presumed to be reflected in scores that concatenate responses in domain-appropriate ways and are generalizable across a range of tasks, settings, and occasions. Furthermore, score-based interpretations and actions are typically extrapolated beyond the test context on the basis of presumed or documented relationships with nontest behaviors and anticipated outcomes or consequences. The challenge in test validation is to link these inferences to convergent evidence supporting them as well as to discriminant evidence discounting plausible rival inferences. Evidence pertinent to all of these aspects needs to be integrated into an overall validity judgment to sustain score inferences and their action implications, or else provide compelling reasons why not, which is what is meant by validity as a unified concept.

The principles of unified validity provide a framework for evaluating all educational and psychological measurement, including language testing and washback. In this connection, it is important to underscore that washback is not simply good or bad teaching or learning practice that might occur with or without the test, but rather good or bad practice that is evidentially linked to the introduction and use of the test. In the context of unified validity, evidence of washback is an instance of the consequential aspect of construct validity, which is only one of six important aspects or forms of evidence contributing to the validity of language test interpretation and use. Valid tests grounded in all six aspects of construct validity, by attempting to minimize construct underrepresentation and construct irrelevancies, should increase the likelihood of positive washback and help to distinguish test washback per se from good and bad educational practices regardless of test quality.

REFERENCES

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*, 115-129.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validation argument. In H. Wainer & H. Braun (Eds.), *Test validity*. Hillsdale, NJ: Erlbaum.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179-197.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105-146). New York: Macmillan.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology*, *10*, 121-131.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, *18*(9), 27-32.

Guilford, J. P. (1936). *Psychometric methods*. New York: McGraw-Hill.

Guion, R. M. (1976). Recruiting, selection, and job placement. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 777-828). Chicago: Rand McNally.

Hunter, J. E., Schmidt, F. L., & Jackson, C. B. (1982). *Advanced meta-analysis: Quantitative methods of cumulating research findings across studies*. San Francisco: Sage.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*, 527-535.

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, *16*, 294-304.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635-694 (Monograph Supplement 9).

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing policy. *Educational Psychologist*, *17*, 67-91.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. London: NFER/Nelson.

Peak, H. (1953). Problems of observation. In L. Festinger & D. Katz (Eds.), *Research methods in the behavioral sciences* (pp. 243-299). Hinsdale, IL: Dryden Press.

Resnick, L. B., & Resnick, D. P. (1991). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Sheingold, K., Heller, J. I., & Paulukonis, S. T. (1994). *Actively seeking evidence: Teacher change through assessment development* (Center for Performance Assessment, MS#94-04). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, *19*, 405-450.

Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research*, *40*, 371-396.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan*, *79*, 703-713.

Wiggins, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappan*, *75*, 200-214.